Description_Capstone_Berlin_Stations_organicFood

June 7, 2021

0.0.1 ++++++ This notebook will mainly be used for the Coursera Capstone Project ++++++

1 Capstone Project Coursera for Data Science: The battle of neighborhoods or the battle of "organic Food & Beverages" at the close vicinity of Berlin metro stations

1.0.1 IBM Data Science Course, attendee Dr.B.Bayer, June 2021

1.1 Introduction

1.1.1 Background Information

Berlin is the capital of Germany with about 3.7 million inhabitants and a city full of movement. On average, every Berliner travels more than three times a day. Local public transport plays a special role in this. Approximately 50 percent of Berlin's households are car-free, and with 324 cars per 1,000 inhabitants, the city has the lowest motorization rate in Germany. In addition, even people who do have a car often use other means of transportation. Public transportation also plays a central role for commuting to work. For example, about 40% of all commuters use public transportation. [1]

Another strongly growing trend is the demand for organic products and therefore incoming, the desire for a healthy diet [2]. In train stations and metro stations, there are already numerous possibilities for food intake, and so numerous small and large stores offer food and drinks to take away. However, these offers are very often not considered healthy at all and thus contradict the desire for a healthy diet. The market for healthy take-away products is still relatively small and this results in a large potential market for investors.

1.1.2 Problem Statement

With the aforementioned prospect, various stakeholders (entrepreneurs, investors) may be interested to explore the organic food and beverage (F&B) shop business opportunities in the very close vicinity of Berlins metro stations. This data science project is thus carried out to help them answer the following question: Which of the Berlins metro stations are strategic for opening an "organic F&B" business?

1.2 Data

In order to explore potential answer to the problem, the following data are required:

- Metro stations of Berlin city with their geographical coordinates [3]. They are required to utilize Foursquare API in the subsequent step.

- Information about venues of the station: the names, category, venue latitudes, venue longitudes. These are obtained using Foursquare API [4]. The stations will be clustered based on their venues to find the best location candidates for "organic F&B".

1.3 Methodology

This section represents the main components of the report. It starts with data extraction (web scraping) of Berlin metro stations and retrieval of geographical coordinates. Leveraging Foursquare API, these coordinates data are given as inputs to explore venues within the stations.

One-hot encoding is performed to analyze and narrow down the most common venues in each of the station. Given all the venues surrounding them, the stations are clustered using K-means algorithm. The number of optimal clusters is decided using the elbow method and silhouette score. Each cluster is separately analyzed to examine one discriminating venue that characterizes them. Analysis of the clusters and visualization will give insights as to where the strategic regions to set up the business.

The following cell contains all the necessary Python libraries.

```python
# import libraries
import numpy as np
import pandas as pd
import requests
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
import json, lxml

from pandas.io.json import json_normalize # tranform JSON file into a pandas
 →dataframe
#from geopy.geocoders import Nominatim # convert an address into latitude and
 →longitude values

from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from bs4 import BeautifulSoup

try:
    import folium
except:
    !pip install folium
    import folium
```

```
try:
    import geopy
except:
    !pip install geopy
    import geopy
from geopy.geocoders import Nominatim

print("Libraries installed.")
```

Libraries installed.